Global Identifiers for the Cancer Bioinformatics Grid

Robert J Robbins Fred Hutchinson Cancer Research Center

Harold Solbrig Mayo Clinic

BACKGROUND ISSUES

Philosophical Background: Identity

- Concept of identity still subject to metaphysical distinctions:
  - NUMERICAL IDENTITY: one thing being the one and only such thing in the universe - e.g., there should be one and only human being associated with a patient ID
  - QUALITATIVE IDENTITY: two things being identical (sufficiently similar) in enough properties to be perfectly interchangeable (for some purpose) – e.g., there are many books associated with an ISBN identifier

Philosophical Background: Properties

- Properties are subject to identity-related distinctions:
  - ACCIDENTAL PROPERTIES: properties of an object that are contingent – that is, properties that are free to change without affecting the identity of the object
  - ESSENTIAL PROPERTIES: non-contingent properties – that is, properties which DEFINE the identity of the object and thus which cannot change without affecting the identity of the object (for some purpose)

Philosophical Background: Properties

- Properties are subject to identity-related distinctions:

Recognizing the distinction between essential and accidental properties will be critical in developing a successful identifier scheme for caBIG.

Especially challenging will be the fact that whether a particular property is essential or not is often context dependent.

s,
ich
ome

Philosophical Background: Properties

- Properties are subject to identity-related distinctions:
  - INTRINSIC PROPERTIES: properties of an object that are properties of the thing itself
  - EXTRINSIC PROPERTIES: properties of the object that are properties of the object’s relationship to other objects external to itself

Philosophical Background: Properties

- Properties are subject to identity-related distinctions:
  - INTRINSIC PROPERTIES: properties of an object that are properties of the thing itself
  - EXTRINSIC PROPERTIES: properties of the object that are properties of the object’s relationship to other objects external to itself

Identifying tandemly duplicated genes is a perfect example of the need to distinguish between extrinsic and intrinsic properties.

Philosophical Background: Identification

- "Identification" is a process that reduces ambiguity. Ambiguity reducing identification can occur in a number of differ ways:
  - INDIVIDUAL SPECIFICATION: denoting an individual object without identifying either its class membership or its individuality - e.g., "this thing"
  - CLASS IDENTIFICATION: specifying than an object is a member of a class of objects that are sufficiently similar that the objects may be considered interchangeable (for some purpose) – e.g., "this book is Darwin's *Origin of Species"*
  - INDIVIDUAL IDENTIFICATION: specifying that an object is in fact a PARTICULAR genuinely unique object in the universe – e.g., this book is Darwin's own personally annotated copy of *Origin of Species*"

Philosophical Background: Identification

- “Identification” is a process that reduces ambiguity. Ambiguity reducing identification can occur in a number of differ ways:

Note that as we move along this continuum our notion of “essential properties” changes.

This shows that the concept of identity can be context dependent.

  - INDIVIDUAL IDENTIFICATION: specifying that an object is in fact a PARTICULAR genuinely unique object in the universe – e.g., this book is Darwin’s own personally annotated copy of *Origin of Species*”

Practical Issues:
Identifying What?

- Digital identifiers (IDs) perform different kinds of identification:
  - REAL-WORLD IDENTIFIER: identifier serves as a digital token representing a real-world (i.e., non-digital) object (e.g., patient ID); this kind of identifier is often used to associated a digital object (bag of properties) with a real-world object
  - DIGITAL IDENTIFIER: identifier serves as a digital token representing a (published?) digital object (e.g., LSID or URL)

Practical Issues:
Identifying What?

- Digital identifiers (IDs) perform different kinds of identification:
  - REAL-WORLD IDENTIFIER: identifier serves as a digital token representing a real-world (i.e., non-digital) object (e.g., patient ID); this kind of identifier is often used to associated a digital object (bag of properties) with a real-world object
  - DIGITAL IDENTIFIER: identifier serves as a digital token representing a (published?) digital object (e.g., LSID or URL)

This distinction can be hard to make:
What does an IP address identify?

Practical Issues: Identification vs Specification

- Digital identifiers (IDs) can truly identify particular objects or they can merely specify singular objects, with no guarantee of what that singular object is:
  - IDENTIFICATION: the same LSID should always return exactly the same (bit for bit) digital object
  - SPECIFICATION: the same URL is not guaranteed to return the same thing twice

Practical Issues:
Identification vs Specification

Note that these two situations really just represent the opposite ends of a continuum:

At one end EVERY property is essential – at the other end NO property is essential.

At both ends, the relationship of identifier to object is clear. In between, this clarity does not exist and contention can and will exist between identifiers and properties (e.g., the same human being could accidentally be assigned two patient IDs, but we could infer identity from the essential properties).

Practical Issues: Identity Claims

- Different methods exist for answering the question whether or not two objects are the same :
  - DEMONSTRATED IDENTITY: the identifiers are the same and the essential properties are the same
  - INFERRED IDENTITY: the identifiers are different but the essential properties are the same
  - INFERRED NON-IDENTITY: the identifiers are the same, but the essential properties are different
  - ASSERTED IDENTITY: the identifiers are the same, but the state of the essential properties are unknown

Practical Issues:
Identity Claims

- Different methods exist for answering the question whether or not two objects are the same :

and the

With checksums, LSIDs are an instance of DEMONSTRATED identity.

Without checksums, LSIDs are an instance of ASSERTED identity.

t the

  - ASSERTED IDENTITY: the identifiers are the same, but the state of the essential properties are unknown

PRACTICAL MATTERS

Scope:
Aspects of an identifier

- Object or Resource itself (+ version)
- Possible attributes (FOSM - Type Tree) (Schema 1)
- Actual attributes (FOSM – Value Tree)
- Representational form (Schema 2)
  - ASN.1, XML, RDF Triples, CSV, ...

Object Identity and Life Science Research – Robbins, 2004

Scope:
What do we need to identify?

- Digital resources
  - **Semi-permanent resources**
    - Protein sequences, Micro array data, Chromatograms
    - Documents and citations
    - “Possible Attributes” (Schema meaning 1)
    - Representational forms (Schema meaning 2)
    - ...
  - **Dynamic resources**
    - Information about people (name, address, ...)
    - Information about specimens (state, location, ...)
    - Information about researchers, institutions, processes, ...
    - Actual Attributes
  - **Results of queries (synthesized information)**

Scope:
What do we need to identify? (continued)

- Non-digital resources
  - Ontological and taxonomic resources
    - Genes, organism taxonomies, diseases, buildings, trees, ...
  - People
  - Organizations
  - Places
  - Devices
  - ...

Scope:
What else do we need to identify?

?

Resource Characteristics

- Digital Resources
  - **Semi-permanant**
    - Static
    - Immutable (?)
    - Long duration
    - Changes are infrequent and easily tracked
  - **Dynamic resources**
    - Dynamic
    - Frequent changes – interest is in current state of resource
    - Duration can be minutes, hours or days
    - Changes are frequent and often not worth tracking

Resource Characteristics (continued)

- Digital Resources (continued)
  - **Synthesized resources**
    - Reproducible if drawn from static information, dynamic otherwise
    - Schema can be dynamic as well (?)
- Non-digital resources
  - Identity is not dependent on digital information
  - Different characteristics and attributes depending upon context and purpose
  - Immutable (in a computational sense)
  - Changes are “ontological” in nature
    - A thing splits into parts
    - One thing is determined to be the same as another
    - Identity is no longer useful

Identification Schemes

- Life Sciences Identifier (LSID)
- ISO 11179
- Lower level schemes
  - DCE UUID (aka. Microsoft GUID)
  - ISO OID
  - DNS
- ?

Identification Schemes
LSID

Format:

URN:LSID:<authority>:<namespace>:<object>[:<revision>]

<authority> - (usually) DNS name

<namespace> - managed by authority

<object> - managed by authority

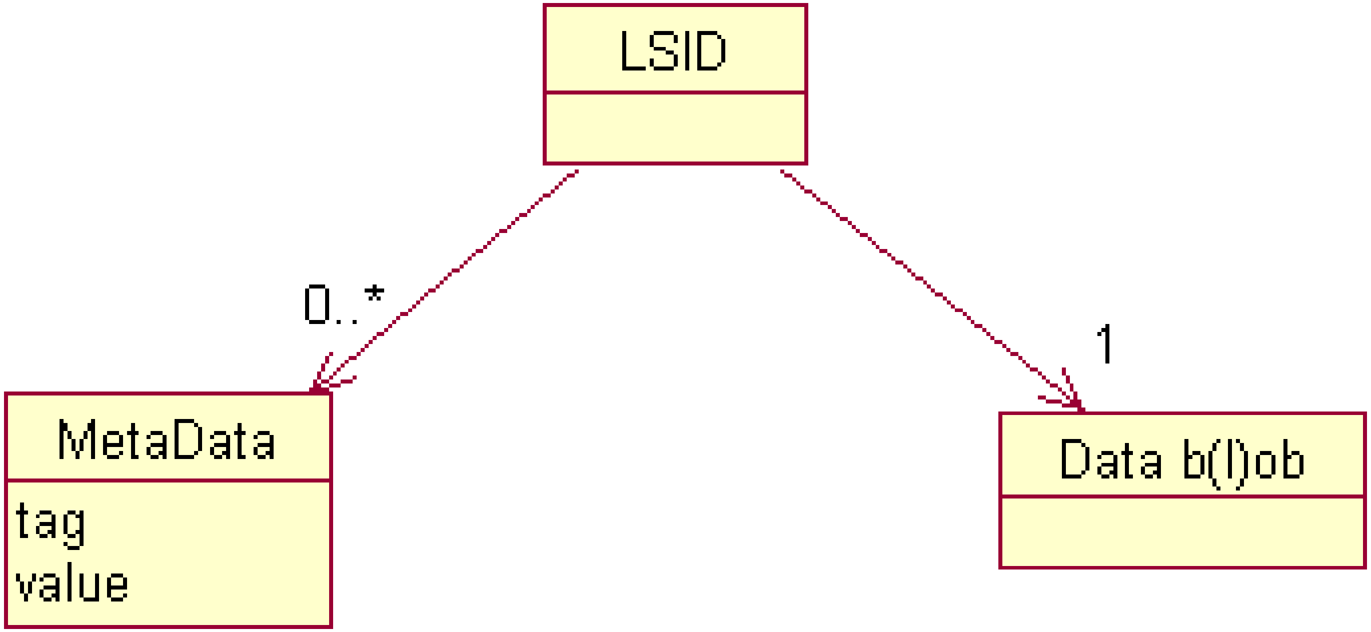
Examples:

URN:LSID:ebi.ac.uk:SWISS-PROT.accession:P34355:3
URN:LSID:rcsb.org:PDB:1D4X:22
URN:LSID:ncbi.nlm.nih.gov:GenBank.accession:NT_001063:2

Identification Schemes
LSID - Scope

This specification addresses the need for a standardized naming schema for *biological entities* in the Life Sciences domains, the need for a service assigning unique identifiers complying with such naming schema, and the need for a resolving service that specifies how to retrieve the entities identified by such naming schema from repositories.

Identification Schemes
LSID - Model



Identification Schemes
LSID – Functional Characteristics

- Resolution Service
- Resolution Discovery Service
- Assigning Service

Identification Schemes
LSID – Synopsis

- Identification of (relatively) static, long-lived resources
- Issues with:
  - Frequent, non-interesting version changes
  - Multiple formats
  - Transformations
    - Relational - projects, joins, restrictions
    - Graph – prune, join, promote
  - Use as foreign keys

Identification Schemes
ISO 11179

Format:

<registration authority identifier> <data identifier> <version>

<registration authority identifier>:

<ICD> <organization id> [<organization part id> [<OPIS>]]

ICD - International Code Identifier (ISO 6253)

OPIS - organization part id source

Example:

0060 ...

Identification Schemes
ISO 11179 - Synopsis

- Not significantly different from LSID from a semantics perspective
- ISO 6523 requirement a significant minus
- No syntax and infrastructure

Identification Schemes
DCE UUID

128 bit binary number

ASCII Format:

hhhhhhhh-hhhh-hhhh-hhhh-hhhhhhhhhhhh

Example:

5d1cb710-1c4b-11d4-bed5-005004b1f42f

Identification Schemes
DCE UUID

- Universally available
- Unlimited supply
- Foundation of Microsoft OLE infrastructure

Identification Schemes
DCE UUID – Synopsis

- Capable of generating identifiers for both static and dynamic resources
- Opaque
  - No problems confusing registrar with location
  - No problems with versioning – versions are external
- Excellent for foreign key purposes
  - fixed size
  - possibility of binary representation
- Already a part of Microsoft infrastructure
- Issues with:
  - Opacity - no way of finding source
    - additional services required for ANY information (database, schema, version, etc.)
  - Network access (vs. LSID)

Identification Schemes
ISO OID

Sequence of arbitrary length (usually small) integers

ASCII Format:

<ra>.<ra>.<ra>....<ra or object>

<ra> - registration authority

Example:

2.16.840.1.113883.6.1

Identification Schemes
ISO OID – Synopsis

- Static focus
- Used in HL7 and other network applications
- Issues with:
  - Non-Opacity - registrar chain often confused with semantics
  - Network access (vs. LSID)
  - Non-numeric values – especially as an object id
  - Database keys
    - arbitrary length
    - dumber databases mistake it for a malformed float

Identification Schemes
Domain Name Service (DNS)

Sequence of dot separated identifiers

ASCII Format:

<ran>...<ra3>.<ra2>.<ra1>

<ra> - registration authority

Example:

informatics.mayo.edu

Identification Schemes
DNS – Synopsis

- Static focus
- Omniscience in WWW
  - Network access and traversal already in place
  - Tooling is everywhere
- Issues with:
  - Non-permanance – names are lost, stolen and reassigned
  - Registration authority hopelessly muddled with semantics
  - Database keys
    - arbitrary length
    - dumber databases mistake it for a malformed float

Discussion Points
Candidates for caBIG identification scheme

- Rule out 11179
  - little advantage over LSID
  - transformation can be specified between 11179 and LSID if necessary
- Rule out ISO OID (Except in case of coding scheme id)
  - little infrastructure support
  - (relatively) unknown outside of HL7 use case
  - Code system identifier registry exists with HL7
    - No need to break it
    - Should specify mapping to LSID

Coding Scheme

urn:lsid:org.iso:2.16.840.1.113883.6.96::

urn:lsid:org.iso:2.16.840.1.113883.6.96:372340004

Concept Code

Discussion Points
Candidates for caBIG identification scheme

- LSID
  - Design intent is static digital information
  - Could possibly be extended to non-digital (conceptual) information
  - Maps identifier to pre-determined set of attributes in fixed schema and representational form
  - Currently doesn't address:
    - Object identifier
    - Schema Meaning 1 (What attributes are possible)
    - Schema Meaning 2 (Representational form)
    - Representational form

Discussion Points
Candidates for caBIG identification scheme

- UUID
  - Ideal choice where opaque identifiers are required
  - Relatively small, fixed size
  - Universally available
  - LSID mapping is possible – one option
    - urn:lsid:org.dce::f81d4fae-7dec-11d0-a765-00a0c91e6bf61:

Not legal currently – no namespace

Discussion Points
Candidates for caBIG identification scheme

- DNS
  - Primary used as an authority identifier
  - To easy to use with URL’s to be generally useful
  - Use should be restricted to component of LSID:

URN:LSID:**ebi.ac.uk**:SWISS-PROT.accession:P34355:3

Discussion Points

- Question: What information do we include in the identifier and what information to we put into a database or service?
  - <Opaque Identifier> →
    - <Object> <ov> <schema> <sv> <rv> + <value>
  - <Object + ov + schema + sv + rv> →
    - <value>
  - ???

Recommendations

- Use separate identifiers for:
  - object
  - schema
  - representational form
  - attribute
  - <object + ov + schema + sv + representational form>
- Version identifiers are
  - Optional – if omitted, “latest” is presumed
  - Opaque – nothing may be presumed from numeric or other collating order

Recommendations (continued)

- **“Object” identifiers**
  - UUID as primary identifier
    - Define formal UUID ← → URN transformation

    URN:UUID:f81d4fae-7dec-11d0-a765-00a0c91e6bf61
  - Syntax should allow an optional display name

    f81d4fae-7dec-11d0-a765-00a0c91e6bf61:John Smith
  - Separate version identifier
    - Service to provide “latest” version for object

    f81d4fae-7dec-11d0-a765-00a0c91e6bf61:John Smith:117
  - LSID map not appropriate because no data is bound

Recommendations (continued)

- **Schema identifiers**
  - Schema format is always XML Schema
  - Primary identifier is LSID URN
    - **Use LSID format as name**

      URN:LSID:org.mayo.edu:patientSchemas:117932:1
    - **Should we spec UUID?**

      URN:LSID:org.dce::f81d4fae-7dec-11d0-a765-00a0c91e6bf61:1

Recommendations (continued)

- **Representational form**
  - Default is XML (need to select version – 1.1?)
  - Identifier is combination of:
    - Mime type
    - URN of form definition

  URN:ISO:2.16.840.1.113883.5.79#text%2Cplain

Recommendations (continued)

- **Attribute**
  - The name of an attribute is an “object” (????)

Recommendations (continued)

- **Exact digital image:**
  <object + ov + schema + sv + representational form>

  - Primary identifier is UUID
  - Transform is available into LSID

    urn:lsid:org.iso::f81d4fae-7dec-11d0-a765-00a0c91e6bf61:

Recommendations Summary

- Object identifiers, schema identifiers and “digital images”:
  - Naming mechanism is UUID within caGRID
  - Use existing mechanisms outside of caGRID
    - GenBank & the like
    - Existing schemas
  - LSID mapping of UUID exists where appropriate
- Representational forms
  - Default is XML
  - Identified by Mime types
  - URN is LSID w/ Mime type OID
- Ontologies, code sets and the like
  - Naming mechanism is ISO OID
  - LSID mapping where appropriate
- Versions
  - Optional – if omitted, “latest” is presumed
  - Opaque – nothing may be presumed from numeric or other collating order

Recommendations

- Additional services as required
  - What representational forms are available for object x <v>?
  - What schemas can be used to represent object x?
  - What schemas contain attribute a?
  - ...